EDAv1 Kelompok 1 2023

September 24, 2023

1 Kelompok 1

1. Meiva Labibah Putri (2204343)
2. Nabil Hanif Achmaddiredja (2205905)
3. Tattha Maharany Yasmin Akbar (2201805)
4. Muhammad Yusdan Ali Batubara (2206847)
5. Ahmad Taufiq Hidayat (2202074)

2 Praproses Data

Memperbaiki atribut

1. Memperbaiki atribut pada setiap kolom dari uppercase menjadi title

Mengubah datatype

1. Mengubah datatype tahun dari int menjadi datetime

Transformasi data

1. Mengubah atribut Makanan pada kolom kategori_usaha menjadi Makanan & Minuman
2. Mengubah atribut Minuman pada kolom kategori_usaha menjadi Makanan & Minuman
3. Mengubah atribut Batik pada kolom kategori_usaha menjadi Fashion
4. Pembatasan rentang tahun pada dataframe Kemiskinan dari tahun 2017 - 2022

Melakukan drop pada atribut

1. Penghapusan atribut yang redundan

[ ]:
```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

2.1 Load Dataset

[ ]:
```
df_umkm = pd.
  ↪read_excel("diskuk-od_17371_jml_ush_mikro_kecil_menengah_umkm__kabupatenkota_kateg_data.
  ↪xlsx")
df_umkm.head()
```

[ ]:

	id	kode_provinsi	nama_provinsi	kode_kabupaten_kota	nama_kabupaten_kota	kategori_usaha	jumlah_umkm	satuan	tahun
0	1	32	JAWA BARAT	3201	KABUPATEN BOGOR	AKSESORIS	927	UNIT	2017
1	2	32	JAWA BARAT	3201	KABUPATEN BOGOR	BATIK	927	UNIT	2017
2	3	32	JAWA BARAT	3201	KABUPATEN BOGOR	BORDIR	132	UNIT	2017
3	4	32	JAWA BARAT	3201	KABUPATEN BOGOR	CRAFT	33111	UNIT	2017
4	5	32	JAWA BARAT	3201	KABUPATEN BOGOR	FASHION	32316	UNIT	2017

[ ]:
```
df_kemiskinan = pd.
↪read_csv("bps-od_20003_angka_garis_kemiskinan_berdasarkan_kabupatenkota_data.
↪csv")
df_kemiskinan.head()
```

[ ]:

	id	kode_provinsi	nama_provinsi	kode_kabupaten_kota	nama_kabupaten_kota	angka_garis_kemiskinan	satuan	tahun
0	1	32	JAWA BARAT	3201	KABUPATEN BOGOR	130927	RUPIAH/KAPITA/BULAN	2004
1	2	32	JAWA BARAT	3202	KABUPATEN SUKABUMI	111202	RUPIAH/KAPITA/BULAN	2004
2	3	32	JAWA BARAT	3203	KABUPATEN CIANJUR	121902	RUPIAH/KAPITA/BULAN	2004
3	4	32	JAWA BARAT	3204	KABUPATEN BANDUNG	133578	RUPIAH/KAPITA/BULAN	2004
4	5	32	JAWA BARAT	3205	KABUPATEN GARUT	108266	RUPIAH/KAPITA/BULAN	2004

2.2 Memperbaiki nama atribut di tiap dataframe

[ ]:
```
df_umkm['nama_provinsi'] = df_umkm['nama_provinsi'].str.title()
df_umkm['nama_kabupaten_kota'] = df_umkm['nama_kabupaten_kota'].str.title()
df_umkm['kategori_usaha'] = df_umkm['kategori_usaha'].str.title()
df_umkm['satuan'] = df_umkm['satuan'].str.title()
df_umkm.head()
```

[ ]:

	id	kode_provinsi	nama_provinsi	kode_kabupaten_kota	nama_kabupaten_kota	kategori_usaha	jumlah_umkm	satuan	tahun
0	1	32	Jawa Barat	3201	Kabupaten Bogor	Aksesoris	927	Unit	2017
1	2	32	Jawa Barat	3201	Kabupaten Bogor	Batik	927	Unit	2017
2	3	32	Jawa Barat	3201	Kabupaten Bogor	Bordir	132	Unit	2017
3	4	32	Jawa Barat	3201	Kabupaten Bogor	Craft	33111	Unit	2017
4	5	32	Jawa Barat	3201	Kabupaten Bogor	Fashion	32316	Unit	2017

```
  kategori_usaha  jumlah_umkm satuan  tahun
0      Aksesoris          927   Unit   2017
1          Batik          927   Unit   2017
2         Bordir          132   Unit   2017
3          Craft        33111   Unit   2017
4        Fashion        32316   Unit   2017
```

[ ]:
```
df_kemiskinan['nama_provinsi'] = df_kemiskinan['nama_provinsi'].str.title()
df_kemiskinan['nama_kabupaten_kota'] = df_kemiskinan['nama_kabupaten_kota'].str.
 ↪title()
df_kemiskinan['satuan'] = df_kemiskinan['satuan'].str.title()
df_kemiskinan.head()
```

[ ]:
```
   id  kode_provinsi nama_provinsi  kode_kabupaten_kota nama_kabupaten_kota  \
0   1             32    Jawa Barat                 3201     Kabupaten Bogor
1   2             32    Jawa Barat                 3202  Kabupaten Sukabumi
2   3             32    Jawa Barat                 3203   Kabupaten Cianjur
3   4             32    Jawa Barat                 3204   Kabupaten Bandung
4   5             32    Jawa Barat                 3205     Kabupaten Garut

   angka_garis_kemiskinan              satuan  tahun
0                  130927  Rupiah/Kapita/Bulan   2004
1                  111202  Rupiah/Kapita/Bulan   2004
2                  121902  Rupiah/Kapita/Bulan   2004
3                  133578  Rupiah/Kapita/Bulan   2004
4                  108266  Rupiah/Kapita/Bulan   2004
```

2.3 Eksplorasi dataframe UMKM dan Kemiskinan

[ ]:
```
df_umkm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1350 entries, 0 to 1349
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   1350 non-null   int64
 1   kode_provinsi        1350 non-null   int64
 2   nama_provinsi        1350 non-null   object
 3   kode_kabupaten_kota  1350 non-null   int64
 4   nama_kabupaten_kota  1350 non-null   object
 5   kategori_usaha       1350 non-null   object
 6   jumlah_umkm          1350 non-null   int64
 7   satuan               1350 non-null   object
 8   tahun                1350 non-null   int64
dtypes: int64(5), object(4)
memory usage: 95.1+ KB
```

[ ]:
```
df_kemiskinan.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 513 entries, 0 to 512
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      513 non-null    int64
 1   kode_provinsi           513 non-null    int64
 2   nama_provinsi           513 non-null    object
 3   kode_kabupaten_kota     513 non-null    int64
 4   nama_kabupaten_kota     513 non-null    object
 5   angka_garis_kemiskinan  513 non-null    int64
 6   satuan                  513 non-null    object
 7   tahun                   513 non-null    int64
dtypes: int64(5), object(3)
memory usage: 32.2+ KB
```

2.3.1 Mengganti datatype atribut

[ ]:
```
df_umkm["tahun"] = pd.to_datetime(df_umkm["tahun"], format="%Y")
df_umkm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1350 entries, 0 to 1349
Data columns (total 9 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   1350 non-null   int64
 1   kode_provinsi        1350 non-null   int64
 2   nama_provinsi        1350 non-null   object
 3   kode_kabupaten_kota  1350 non-null   int64
 4   nama_kabupaten_kota  1350 non-null   object
 5   kategori_usaha       1350 non-null   object
 6   jumlah_umkm          1350 non-null   int64
 7   satuan               1350 non-null   object
 8   tahun                1350 non-null   datetime64[ns]
dtypes: datetime64[ns](1), int64(4), object(4)
memory usage: 95.1+ KB
```

[ ]:
```
df_kemiskinan["tahun"] = pd.to_datetime(df_kemiskinan["tahun"], format="%Y")
df_kemiskinan.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 513 entries, 0 to 512
Data columns (total 8 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      513 non-null    int64
```

```
 1   kode_provinsi           513 non-null    int64
 2   nama_provinsi           513 non-null    object
 3   kode_kabupaten_kota     513 non-null    int64
 4   nama_kabupaten_kota     513 non-null    object
 5   angka_garis_kemiskinan  513 non-null    int64
 6   satuan                  513 non-null    object
 7   tahun                   513 non-null    datetime64[ns]
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 32.2+ KB
```

2.4 Transformasi Data UMKM

[ ]:
```
df_umkm.kategori_usaha = df_umkm.kategori_usaha.replace({"Makanan": "Makanan &␣
↪Minuman", "Minuman": "Makanan & Minuman", "Batik": "Fashion"})
```

2.4.1 Penggabungan kategori usaha pada dataframe UMKM

[ ]:
```
df_umkm = df_umkm.groupby(['nama_provinsi', 'kode_kabupaten_kota',␣
↪'nama_kabupaten_kota', 'kategori_usaha', 'satuan', 'tahun'])['jumlah_umkm'].
↪sum().reset_index()
df_umkm = df_umkm[['nama_provinsi', 'kode_kabupaten_kota',␣
↪'nama_kabupaten_kota', 'kategori_usaha','jumlah_umkm', 'satuan', 'tahun']]
df_umkm.head()
```

[ ]:
```
  nama_provinsi  kode_kabupaten_kota nama_kabupaten_kota kategori_usaha  \
0    Jawa Barat                 3201     Kabupaten Bogor      Aksesoris
1    Jawa Barat                 3201     Kabupaten Bogor      Aksesoris
2    Jawa Barat                 3201     Kabupaten Bogor      Aksesoris
3    Jawa Barat                 3201     Kabupaten Bogor      Aksesoris
4    Jawa Barat                 3201     Kabupaten Bogor      Aksesoris

   jumlah_umkm satuan      tahun
0          927   Unit 2017-01-01
1          984   Unit 2018-01-01
2         1045   Unit 2019-01-01
3         1110   Unit 2020-01-01
4         1179   Unit 2021-01-01
```

2.4.2 Pembatasan DataFrame Kemiskinan dari tahun 2017 sampai 2022

[ ]:
```
df_kemiskinan = df_kemiskinan[(df_kemiskinan['tahun']>'2016-01-01') &␣
↪(df_kemiskinan['tahun']<'2022-01-01')]
df_kemiskinan = df_kemiskinan[['kode_provinsi', 'nama_provinsi',␣
↪'kode_kabupaten_kota', 'nama_kabupaten_kota', 'angka_garis_kemiskinan',␣
↪'satuan', 'tahun']]
```

2.5 Merge Kedua DataFrame

```
[ ]: df2 = pd.merge(df_umkm, df_kemiskinan,how="left", on=["kode_kabupaten_kota",␣
      ↪"tahun"])
     df2.head()
```

```
[ ]:   nama_provinsi_x  kode_kabupaten_kota nama_kabupaten_kota_x kategori_usaha  \
     0      Jawa Barat                 3201       Kabupaten Bogor      Aksesoris
     1      Jawa Barat                 3201       Kabupaten Bogor      Aksesoris
     2      Jawa Barat                 3201       Kabupaten Bogor      Aksesoris
     3      Jawa Barat                 3201       Kabupaten Bogor      Aksesoris
     4      Jawa Barat                 3201       Kabupaten Bogor      Aksesoris

        jumlah_umkm satuan_x      tahun  kode_provinsi nama_provinsi_y  \
     0          927     Unit 2017-01-01             32      Jawa Barat
     1          984     Unit 2018-01-01             32      Jawa Barat
     2         1045     Unit 2019-01-01             32      Jawa Barat
     3         1110     Unit 2020-01-01             32      Jawa Barat
     4         1179     Unit 2021-01-01             32      Jawa Barat

       nama_kabupaten_kota_y  angka_garis_kemiskinan             satuan_y
     0       Kabupaten Bogor                  337550  Rupiah/Kapita/Bulan
     1       Kabupaten Bogor                  359787  Rupiah/Kapita/Bulan
     2       Kabupaten Bogor                  373799  Rupiah/Kapita/Bulan
     3       Kabupaten Bogor                  402877  Rupiah/Kapita/Bulan
     4       Kabupaten Bogor                  418483  Rupiah/Kapita/Bulan
```

2.5.1 Penghapusan nama atribut yang terduplikasi

```
[ ]: df2 = df2.drop(columns=['nama_provinsi_y', 'nama_kabupaten_kota_y', 'satuan_x',␣
      ↪'satuan_y'])
     df2 = df2.rename(columns={'nama_provinsi_x': 'nama_provinsi',␣
      ↪'nama_kabupaten_kota_x': 'nama_kabupaten_kota'})
     df2["kode_provinsi"] = df2["kode_provinsi"].astype("category")
     df2["kode_kabupaten_kota"] = df2["kode_kabupaten_kota"].astype("category")
     df2.head()
```

```
[ ]:   nama_provinsi kode_kabupaten_kota nama_kabupaten_kota kategori_usaha  \
     0    Jawa Barat                3201     Kabupaten Bogor      Aksesoris
     1    Jawa Barat                3201     Kabupaten Bogor      Aksesoris
     2    Jawa Barat                3201     Kabupaten Bogor      Aksesoris
     3    Jawa Barat                3201     Kabupaten Bogor      Aksesoris
     4    Jawa Barat                3201     Kabupaten Bogor      Aksesoris

        jumlah_umkm      tahun kode_provinsi  angka_garis_kemiskinan
     0          927 2017-01-01            32                  337550
     1          984 2018-01-01            32                  359787
```

```
2          1045 2019-01-01                32                  373799
3          1110 2020-01-01                32                  402877
4          1179 2021-01-01                32                  418483
```

[ ]:
```
df2 = df2[['kode_provinsi', 'nama_provinsi', 'kode_kabupaten_kota',␣
 ↪'nama_kabupaten_kota', 'kategori_usaha', 'jumlah_umkm',␣
 ↪'angka_garis_kemiskinan', 'tahun']]
df2.head()
```

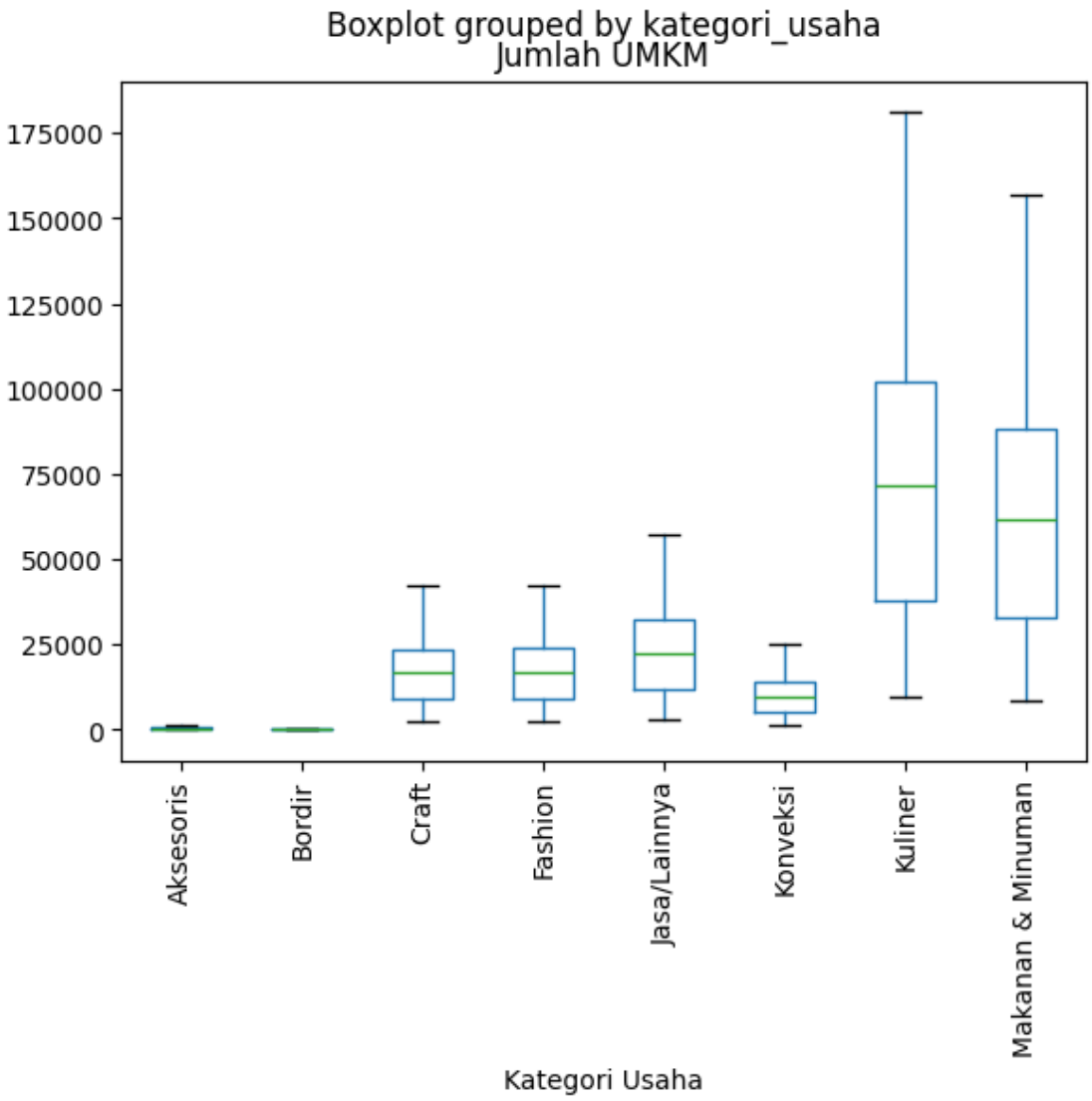
[ ]:
```
  kode_provinsi nama_provinsi kode_kabupaten_kota nama_kabupaten_kota  \
0            32    Jawa Barat                3201     Kabupaten Bogor
1            32    Jawa Barat                3201     Kabupaten Bogor
2            32    Jawa Barat                3201     Kabupaten Bogor
3            32    Jawa Barat                3201     Kabupaten Bogor
4            32    Jawa Barat                3201     Kabupaten Bogor

  kategori_usaha  jumlah_umkm  angka_garis_kemiskinan      tahun
0      Aksesoris          927                  337550 2017-01-01
1      Aksesoris          984                  359787 2018-01-01
2      Aksesoris         1045                  373799 2019-01-01
3      Aksesoris         1110                  402877 2020-01-01
4      Aksesoris         1179                  418483 2021-01-01
```

3 Visualisasi

[ ]:
```
df3 = df2.groupby(['nama_kabupaten_kota', 'angka_garis_kemiskinan',␣
 ↪'tahun'])['jumlah_umkm'].sum().reset_index()
df2017 = df3[df3['tahun'] == '2017-01-01']
df2018 = df3[df3['tahun'] == '2018-01-01']
df2019 = df3[df3['tahun'] == '2019-01-01']
df2020 = df3[df3['tahun'] == '2020-01-01']
df2021 = df3[df3['tahun'] == '2021-01-01']
```

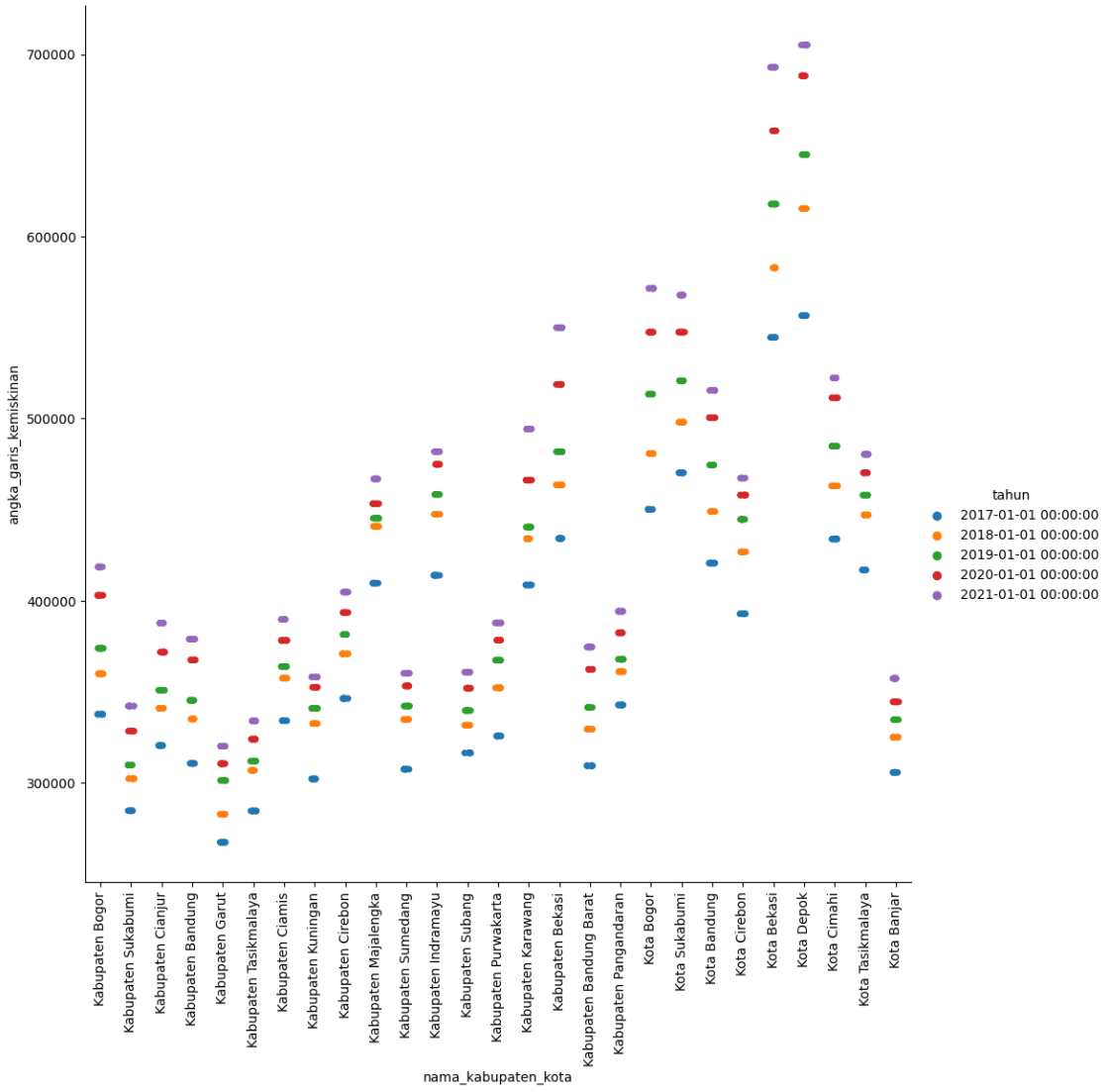
Boxplot untuk menampilkan atribut kategori usaha berdasarkan jumlah umkm

[ ]:
```
df2.boxplot(by ='kategori_usaha', column =['jumlah_umkm'], grid = False)
plt.title("Jumlah UMKM")
plt.xlabel('Kategori Usaha')
plt.xticks(rotation=90)
plt.show()
```



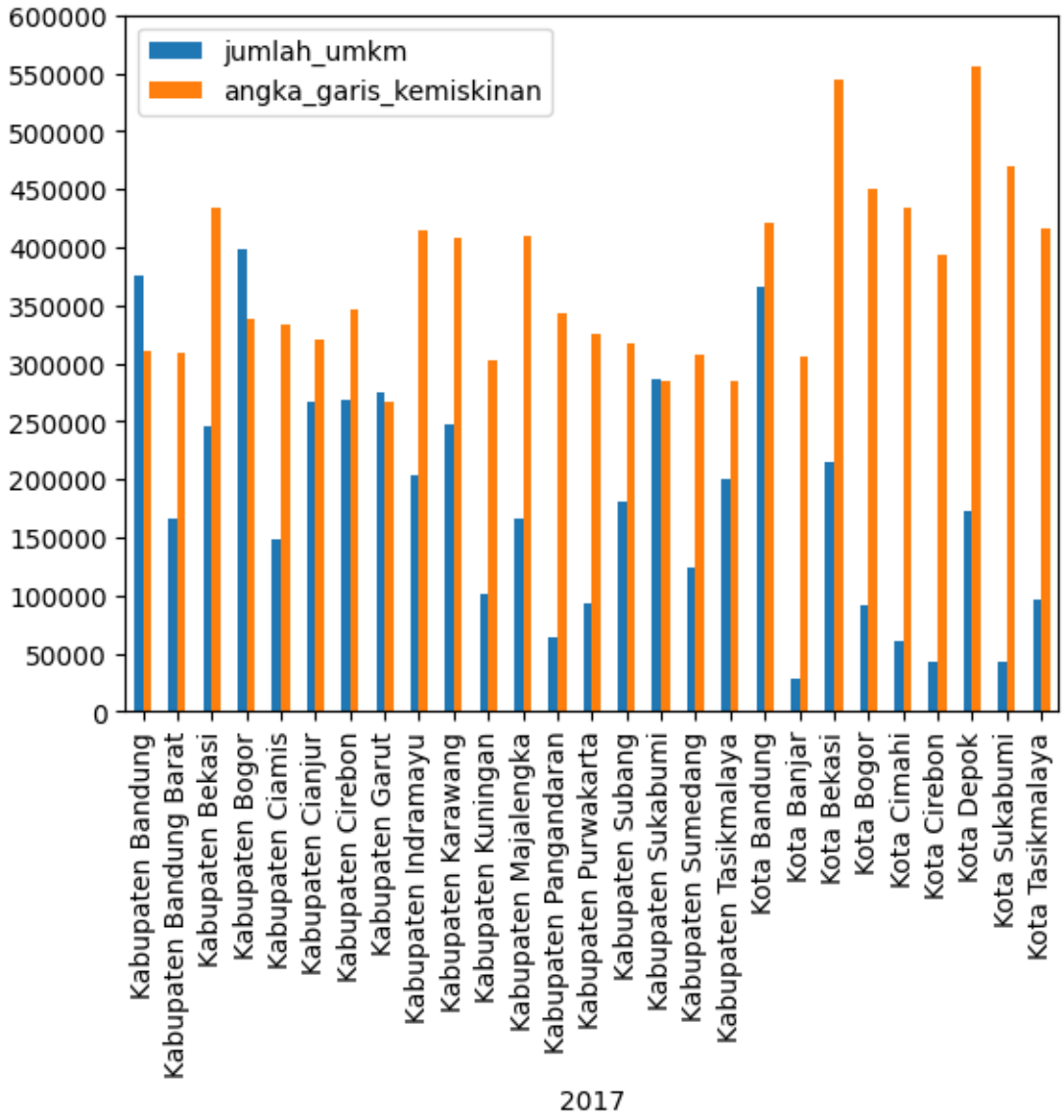
Catplot untuk menampilkan angka kemiskinan pada setiap daerah per tahunnya

[ ]:
```
import warnings
warnings.filterwarnings('ignore')
sns.catplot(x="nama_kabupaten_kota", y="angka_garis_kemiskinan",hue="tahun",␣
↪data=df2, height=10)
plt.xticks(rotation=90)
plt.show()
```



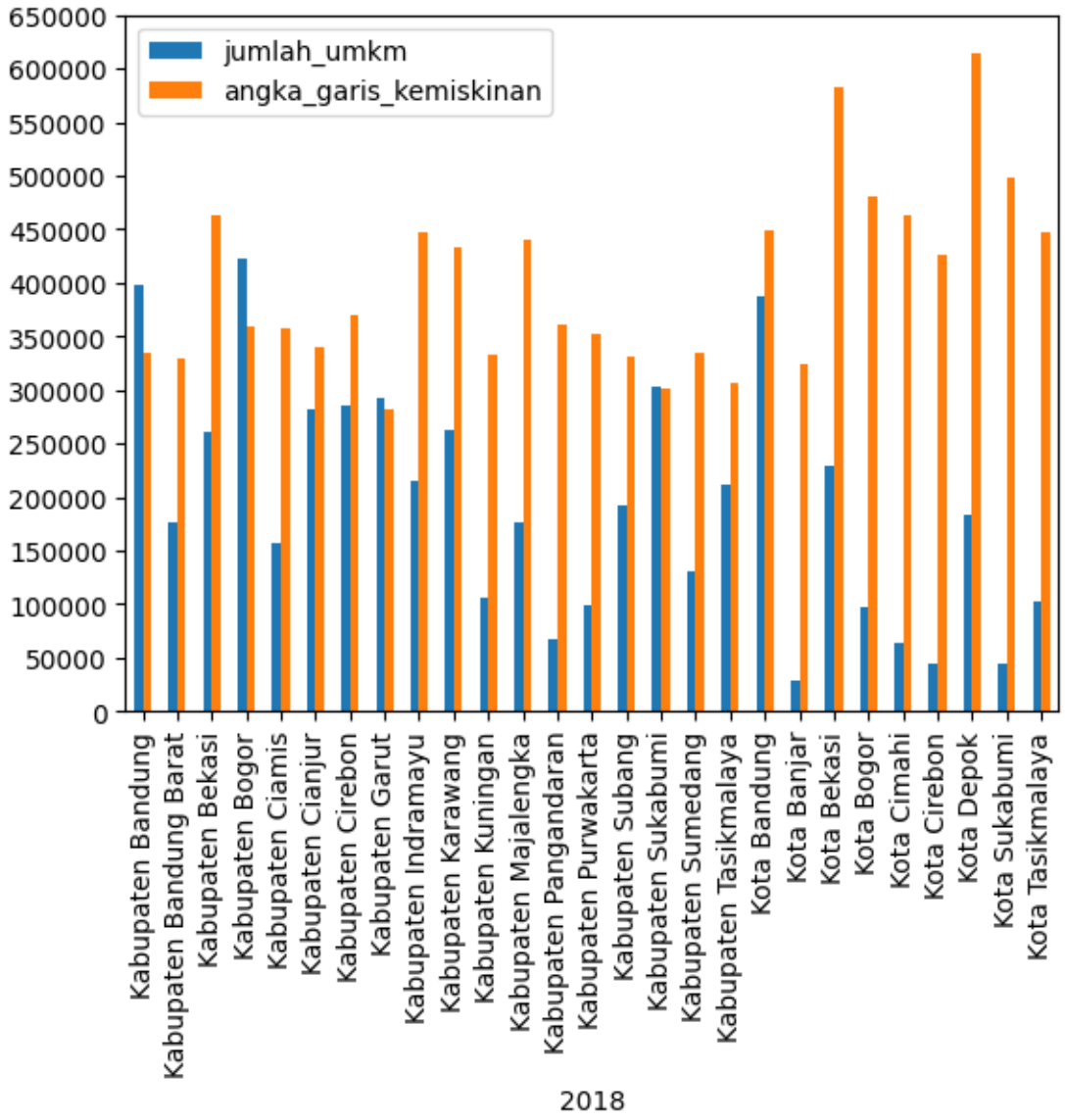
Membandingkan rata-rata jumlah UMKM dengan angka kemiskinan pada tahun 2017

```
[ ]: df_groups2017 = df2017.groupby(['nama_kabupaten_kota'])[['jumlah_umkm',␣
      ↪'angka_garis_kemiskinan']].sum()
     df_groups2017.plot(kind='bar')
     plt.xlabel('2017')
     max_y = np.ceil(df_groups2017[['jumlah_umkm', 'angka_garis_kemiskinan']].max().
      ↪max() / 50000) * 50000
     plt.yticks(np.arange(0, max_y + 1, 50000))
     plt.show()
```



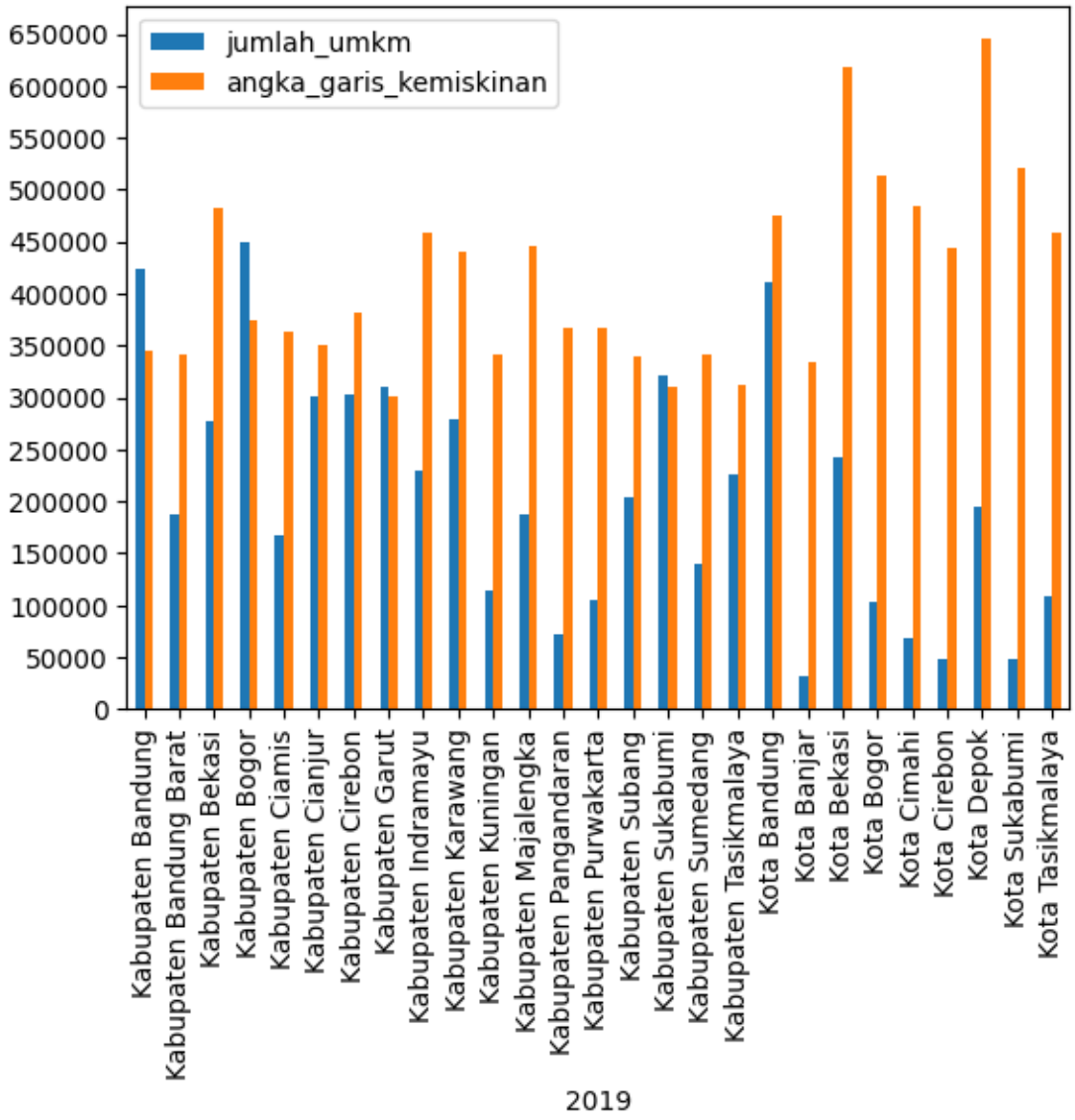
Membandingkan rata-rata jumlah UMKM dengan angka kemiskinan pada tahun 2018

[ ]:
```
df_groups2018 = df2018.groupby(['nama_kabupaten_kota'])[['jumlah_umkm',␣
  ↪'angka_garis_kemiskinan']].sum()
df_groups2018.plot(kind='bar')
plt.xlabel('2018')
max_y = np.ceil(df_groups2018[['jumlah_umkm', 'angka_garis_kemiskinan']].max().
  ↪max() / 50000) * 50000
plt.yticks(np.arange(0, max_y + 1, 50000))
plt.show()
```



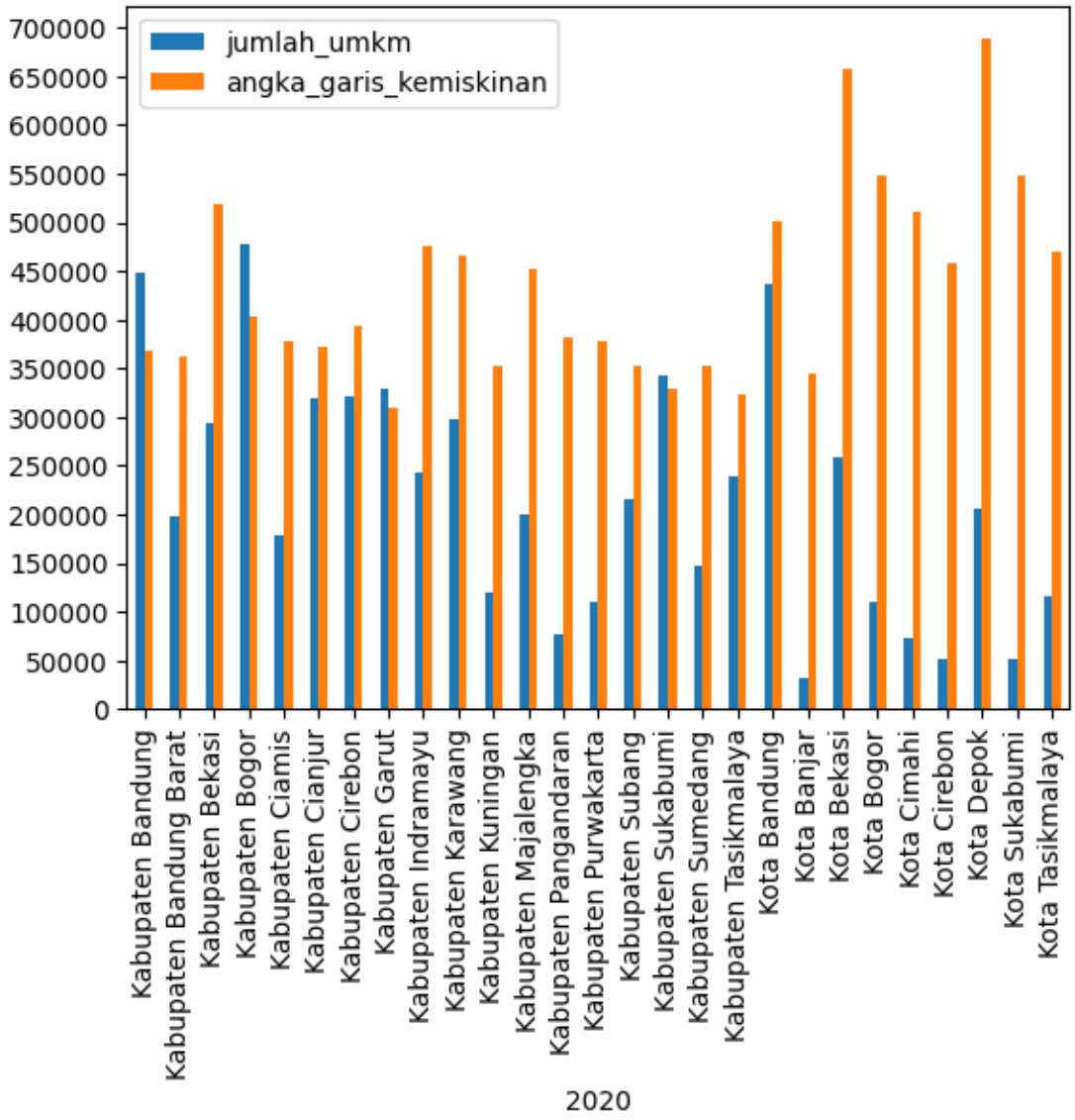
Membandingkan rata-rata jumlah UMKM dengan angka kemiskinan pada tahun 2019

[ ]:
```
df_groups2019 = df2019.groupby(['nama_kabupaten_kota'])[['jumlah_umkm',␣
  ↪'angka_garis_kemiskinan']].sum()
df_groups2019.plot(kind='bar')
plt.xlabel('2019')
max_y = np.ceil(df_groups2019[['jumlah_umkm', 'angka_garis_kemiskinan']].max().
  ↪max() / 50000) * 50000
plt.yticks(np.arange(0, max_y + 1, 50000))
plt.show()
```



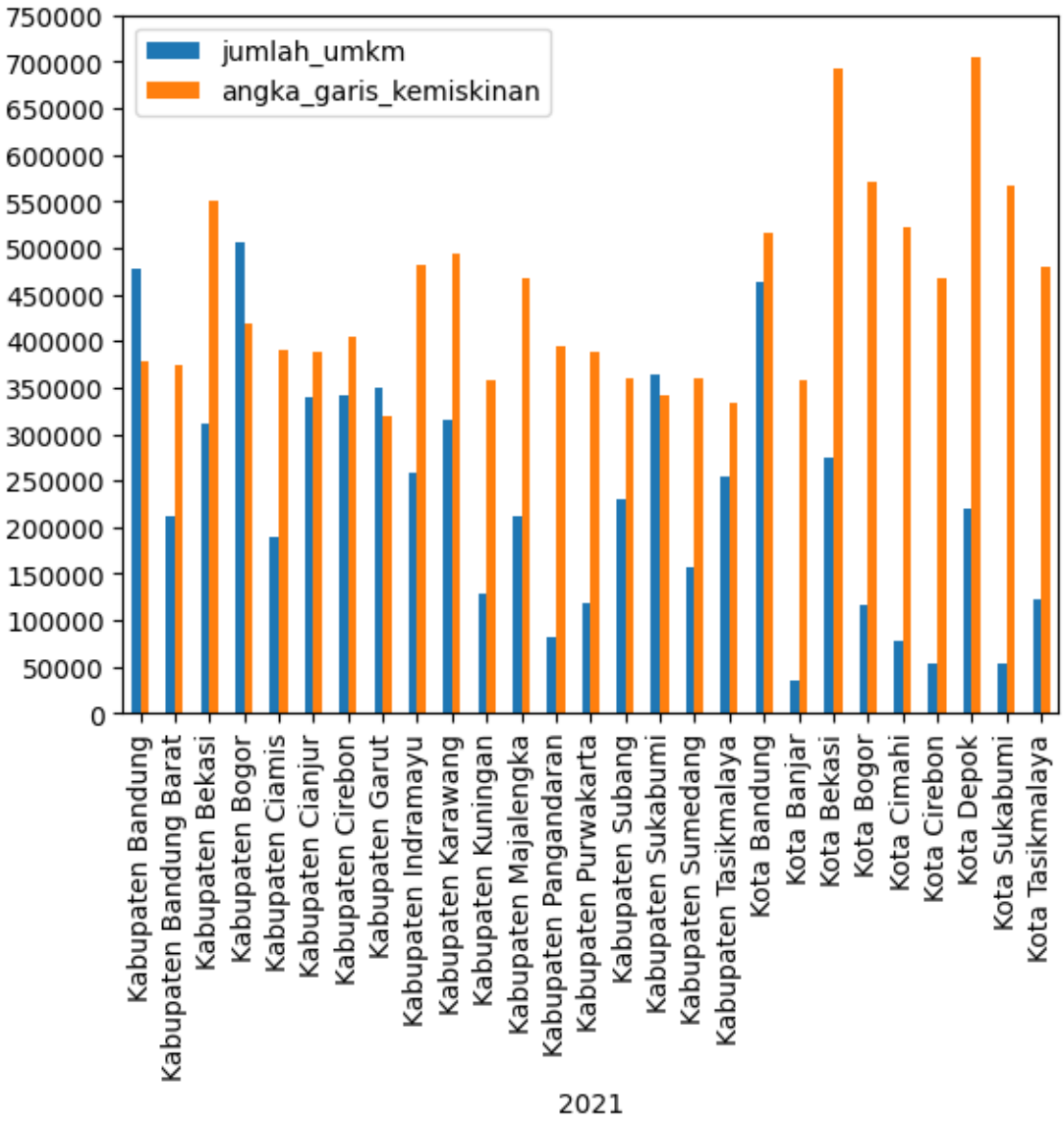
Membandingkan rata-rata jumlah UMKM dengan angka kemiskinan pada tahun 2020

```
[ ]: df_groups2020 = df2020.groupby(['nama_kabupaten_kota'])[['jumlah_umkm',␣
      ↪'angka_garis_kemiskinan']].sum()
     df_groups2020.plot(kind='bar')
     plt.xlabel('2020')
     max_y = np.ceil(df_groups2020[['jumlah_umkm', 'angka_garis_kemiskinan']].max().
      ↪max() / 50000) * 50000
     plt.yticks(np.arange(0, max_y + 1, 50000))
     plt.show()
```



Membandingkan rata-rata jumlah UMKM dengan angka kemiskinan pada tahun 2021

[ ]:
```
df_groups2021 = df2021.groupby(['nama_kabupaten_kota'])[['jumlah_umkm',␣
  ↪'angka_garis_kemiskinan']].sum()
df_groups2021.plot(kind='bar')
plt.xlabel('2021')
max_y = np.ceil(df_groups2021[['jumlah_umkm', 'angka_garis_kemiskinan']].max().
  ↪max() / 50000) * 50000
plt.yticks(np.arange(0, max_y + 1, 50000))
plt.show()
```


jumlah_umkm
angka_garis_kemiskinan
750000
700000
650000
600000
550000
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
Kabupaten Bandung
Kabupaten Bandung Barat
Kabupaten Bekasi
Kabupaten Bogor
Kabupaten Ciamis
Kabupaten Cianjur
Kabupaten Cirebon
Kabupaten Garut
Kabupaten Indramayu
Kabupaten Karawang
Kabupaten Kuningan
Kabupaten Majalengka
Kabupaten Pangandaran
Kabupaten Purwakarta
Kabupaten Subang
Kabupaten Sukabumi
Kabupaten Sumedang
Kabupaten Tasikmalaya
Kota Bandung
Kota Banjar
Kota Bekasi
Kota Bogor
Kota Cimahi
Kota Cirebon
Kota Depok
Kota Sukabumi
Kota Tasikmalaya
2021